Extracting Co-referential Objects Automatically from Multi-source POI Datasets Based on Position-correction and Semantic Matching

Liu Jiping, Wang Yong, Luo An, Dong Chun

Chinese Academy of Surveying and Mapping, No.28 Lianhuachi West Road, Haidian District, Beijing, 100830, China

Abstract. The problem of automatically extracting co-referential objects from multi-source POI datasets in chinese has not been perfectly resolved for the low rate of recall and accuracy caused by the position deviation, names and addresses confusion. This paper proposes an automatic extracting method for co-referential objects from multi-source POI datasets based on position correction and semantic matching. The main step of the method including: (1) semantically selecting entities of same name and same address, (2) locally-position-correction to to-be-merged POI dataset based on triangulation, and (3) extracting remaining co-referential entities by Nearest Neighbor Method and semantic-comparation. Results of experimental tests show that this method achieved good effect in recall and precision rates.

Keywords: Multi-source POI, Co-referential Object, Position-correction, Address standardization, Semantic matching

1. Introduction

The accuracy and currency of POI is crucial for the availability of geographic information services, which requires the database of POI enriching and updating continually. At present, on one hand, POI data updating mainly uses the way of artificial regeneration and the greater the amount of data, the higher the maintenance costs; on the other hand, the POI information resource from the Internet is increasingly rich and can serve as an important data source of POI data updating. Therefore, it is necessary to study the fusion and update of multi-source POI data automatically.

Extracting co-referential object (the same geographic target in real life) automatically is the basis of multi-source POI data fusion. Lacking of unified

entity Naming rules, address coding rules, attribute naming rules and others, combined with the inevitable errors producing when different service providers acquire and process positional information of POI, resulting in the POI data from different sources is often difficult to fuse directly. Mainly displays in the following respects:

- The positioning coordinate of co-referential object in different POI data exists nonlinear deviation and unable to set position together.
- The naming method of co-referential object in different POI data is not identical.
- The address representation of co-referential object in different POI data is not identical.

Therefore, this paper proposes a method of extracting co-referential objects automatically from multi-source POI Datasets based on the position correction and semantic matching. It can solve the low automatic matching rate and accuracy of co-referential object caused by the difference of position, name and address.

2. Related technologies

2.1. Semantic matching of entity's name

The aim of semantic matching of entity's name is to calculate the similarity of entities' name. Now the technology of name matching in english is already relatively mature; but for chinese, this problem is still not completely resolved for structural instability, complex nested relationships, lacking of iconic words etc. At now, the matching method of entity's name in chinese is mainly focus on the matching of the key words contained in names. Liu X proposed a segmentation and matching method based on the lucene Chinese POI name, realizing fuzzy matching according to the different roles of the POI segmentation unit. Zhang X achieved the automatic identification of Chinese organization name by analyzing the structure of the Chinese organization name. Li J presented a method of identifying the Chinese organization name based on template matching according to the unknown words in the Chinese organization name. Yu H proposed a method of Chinese organization name identification based on role labeling.

2.2. Geocoding

Geocoding is a process of associating address information expressing spatial location with space actual coordinates, indicating that making the address data mapped into geographical coordinates. It is a space positioning technology based on the text information of address. The general method is:

Splitting the address string of geocoding, then standardizing address by address model expression, next matching the field value of standardized key address and the corresponding field attribute value of geographical entity in the spatial reference data, finally selecting geographic coordinates of match results and assigning the value to corresponding attribute, so as to realize the effective coding to address. So far, geocoding software tools for the language of english has matured, but those for the language of chinese are still at the preliminary and exploratory stage.

2.3. Position-correction

The main method of Position-correction is to establish the mapping relationship of corresponding point in two different spaces. Conversion and mapping can be achieved between two different coordinate systems, regular deformation error can be avoided or weaken in the acquisition process, and provided technical support for multi-source data. The common transformation method of space position of POI data is: establishing the coordinate transformation model of affine, similarity and projection (like the Remote Sensing Image Correction in Photogrammetry), automatically selecting series point of same name in different data sources and different coordinate systems, calculating the parameters of transformation model using the least squares method, then executing geometric transformation to the maps data.

2.4. Matching of geographic entity

Geographic entity matching aims to identify the same feature in two different data sets. There are many elements for the matching such as location, shape, structure, topology, names, attributes and so on. According to the different elements, the matching method can be divided into geometric matching, topological matching and semantic matching. Geometric matching select candidate entity based on entity space attribute through the calculation of space distance. Topology matching is based on the topological relations measurable and computing of the name and address for the candidate entities. Semantic matching is the method by comparing semantic information of candidate entities.

3. The automatic extracting method of multi-source POI co-referential object

This paper presents an automatic extracting method of Multi-source POI co-referential object with fusing of semantic analysis, geocoding and adjustment processing method. It is shown *Figure 1* and can be described as follows: Firstly, selecting entity of same name and address through address

standardization and similarity matching. Secondly, executing the position correction on the base of the POI set with the same name and address, then integrated computing the corrected data based on the spatial distance and semantic similarity. Finally extracting the co-referential objects from the POI datasets.

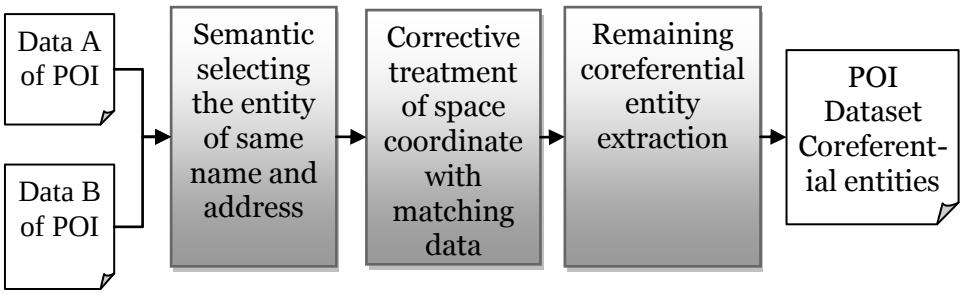

Figure 1. The workflow of extracting multi-source POI co-referential object.

3.1. Semantically selecting the entity of same name and address

3.1.1 Address standardization and semantic matching

Learning from other geocoding model such as DIME model, TIGER model and ESRI model, this paper designed a hierarchical model of geocoding with combing the chinese specifications (*Figure 2*). It mainly consists three parts: administrative division, address and sub-address. Administrative division includes country name, province name, city name, area name and county name. Address includes fundamental and extension part of address. And sub address includes fundamental and extension of sub-address.

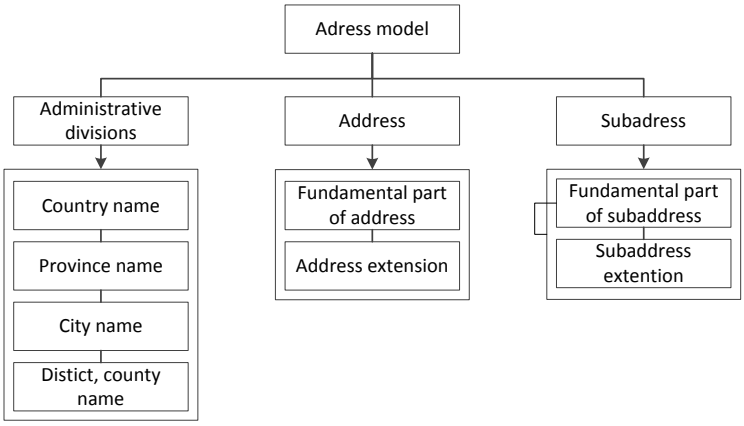

Figure 2. Chinese geocoding mode.

According to the designed model of geocoding, the expression of standard address is:

<The standard address> :: = <[country name] [province name] [city name] [district county name]><[fundamental address] [extended address]><[fundamental sub-address] [extended sub-address]>

The process of address standardization is to express any address based on the mode of standardized address, and automatically fill the missing part of higher level according to the affiliation elements. So that it can form a standardized address expressions with the end of the most low address.

The standardized address matching is executed layer by layer from top to down and high to low. And the semantic matching results can be divided into three cases:

- Exactly matching: the number and name of element layers are completely the same after standardized. This belongs to exactly matching.
- Compatibly matching: when all elements in each layer have matched successfully, another standardized address still exist subordinates address elements. It illustrates the accuracy of the two addresses are different, and they are mutually compatible. This belongs to compatible matching.
- Mismatching: standardized address appears the address elements cannot be matched exactly layer by layer, and it belongs to mismatching.

3.1.2 POI's name Matching with Similarity Calculation

POI's name Matching is a processed to determine whether it is same in name and address according to the text name of POI. In the chinese name expression of POI, it can play different role due to the different location, so the effect in word matching of POI's name is different.

Therefore, we propose a matching method of POI's name based on role tagging which is shown in *Figure 3*. Firstly, tagging the role for words in POI's name with the phrase segmentation and using word dictionary. Secondly extracting the central word in POI's name, then cutting in the POI's name with the help of central word. And finally calculating the similarity of the POI's name with the results of cutting.

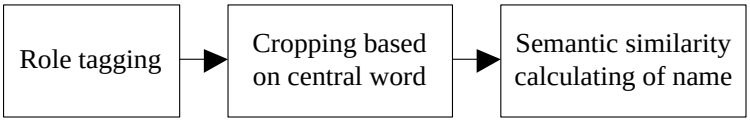

Figure 3. The process of POI's name matching.

The similarity calculation formula of POI's name matching is as follows:

$$P(a,b) = \lambda \cdot P(w_{a-center}, w_{b-center}) + \beta_1 \cdot P(w_{a-1}, w_{b-1}) + \beta_2 \cdot P(w_{a-2}, w_{b-2}) + \ldots + \beta_i \cdot P(w_{a-i}, w_{b-i}) \quad (1)$$

$P(a,b)$ indicates the probability of a and b are the same POI, $w_{a-center}$indicates the central word of a, w_{a-i} indicates the i-th modifier of a, λ , β_1 … β_i indicates the parameter between [0,1] and $\sum(\lambda, \beta_1 \ldots \beta_i) = 1$.

According to the result of name matching, we can divide it into three cases:

- Exactly matching: the center words and all modifiers of POI name are exactly the same, or the center word is the same and most of the modifiers are the same and which make the matching similarity greater than a specific threshold. It belongs to complete matching.
- Type matching: the center word of POI name is the same or belongs to the same class and most of the modifiers are the same. And it makes the matching similarity beyond a specific threshold. It belongs to type matching.
- Mismatching: the center word of POI name does not belong to the same class and most qualifiers are different, and the matching similarity is lower than a specific threshold. It belongs to mismatch.

3.1.3 Selection of POI object with the same name and address

We can make a permutation and combination according to the results of POI address standardization in *Section 3.1.1* and POI's name matching in *Section 3.1.2*, then form a matching result matrix of POI's name and address, as shown in *Table 1*.

name \ Address	Exactly matching	Compatibly matching	Mismatching
Exactly matching	M11	M12	M13
Type matching	M21	M22	M23
Mismatching	M31	M32	M33

Table 1. The matching results of POI object name and address.

- M11 indicates the object set of which both the name and the address of POI are exactly matched.

- M12 indicates the object set of which the name is exactly matched and the address is compatibly matched.
- M13 indicates the object set of which the name is exactly matched and the address is not matched.
- M21 indicates the object set of which the name belongs to type matching and the address is exactly matched.
- M22 indicates the object set of which the name belongs to type matching and the address belongs to compatible matching.
- M23 indicates the object set of which the name belongs to type matching but the address is not matched.
- M31 indicates the object set of which the name is not matched but the address is compatibly matched.
- M32 indicates the object set of which the name is not matched but the address belongs to compatible matching.
- M33 indicates the object set of which neither the name nor the address of POI is matched.

The selection of POI objects with the same name is based on the combined results of name matching and addresses matching, and the POI object in pairs in M11 set as the entities with same name and address can be selected.

3.2. Position-correction

Deviation inevitably exists among different POI datasets, so it is necessary to do position-correction before we start fusing work to reduce the positional deviation between the two sets of POI data. In most cases, the position deviation of different POI data is nonlinear, so we can't use a linear transformation function to carry out simple conversion. This paper presents a processing method of making local linear transformations based on triangulations.

In *Section* 3.1, entities with the same name and same address(Set M) have been extracted. Now, we make Delaunay triangulation to the spatial extent of set B, using the entities in Set M as the verte, forming triangular meshes T (Set T) as following (Figure 4).

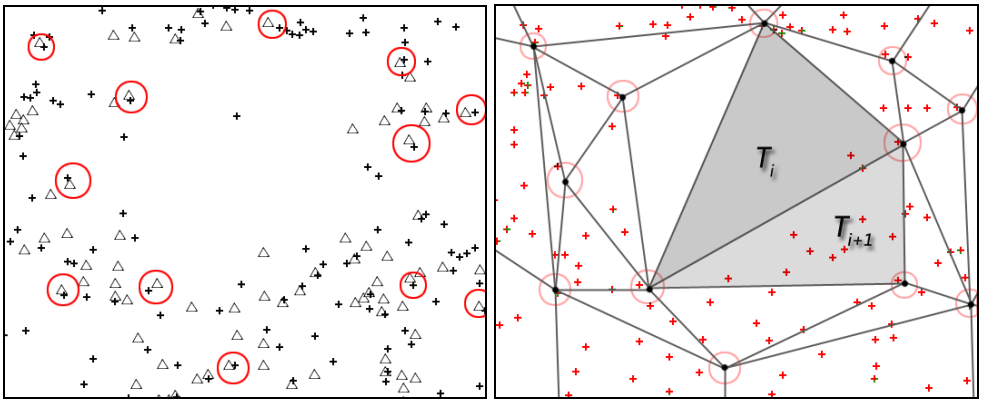

Figure 4. Triangulation of POI set B using the points in Set M

In triangle T_i in Set T, assuming the coordinate transformation of B→A as a local linear transformation and the transformation parameter is $D_i = (a_i, b_i, m_i, n_i)$, the corrective function is given as following :

$$X_a = a_i X_b + b_i \qquad (2)$$

$$Y_a = m_i Y_b + n_i \qquad (3)$$

To get the transform parameters of $D_i = (a_i, b_i, m_i, n_i)$, the method of LS(Least Squares) is used. So the position of the corresponding object in POI set A and B of T_i's three vertices needs to satisfy the following conditions:

$$M(a_i, b_i) = \sum_{k=0}^{n} (X_a - a_i X_b - b_i)^2 = \min \qquad (4)$$

$$M(m_i, n_i) = \sum_{k=0}^{n} (Y_a - m_i Y_b - n_i)^2 = \min \qquad (5)$$

The calculating formula of a_i and b_i is:

$$\begin{cases} (\sum_{k=0}^{n} X_{bk}^2) a_i + (\sum_{k=0}^{n} X_{bk}) b_i = \sum_{k=0}^{n} X_{bk} X_{ak} \\ (\sum_{k=0}^{n} X_{bk}) a_i + (n+1) b_i = \sum_{k=0}^{n} X_{ak} \end{cases} \qquad (6)$$

The calculating formula of m_i and n_i is:

$$\begin{cases} (\sum_{k=0}^{n} Y_{bk}^2)m_i + (\sum_{k=0}^{n} Y_{bk})n_i = \sum_{k=0}^{n} Y_{bk}Y_{ak} \\ (\sum_{k=0}^{n} Y_{bk})m_i + (n+1)n_i = \sum_{k=0}^{n} Y_{ak} \end{cases} \quad (7)$$

When getting the local transformation parameters for each triangle in set T, each triangle (Ti ∈T) is enumerated, within which the position of each POI object will be recalculated using equation (2), (3), to get the new POI set Bt. Pseudo codes in C# style are listed in Table 2:

```
...
ArrayList<POI> Bt;
for each（Triangle t in T）{
        for each ( POI p in t ){
            POI pn = new POI();
            Pn.X = a_i ×P.X+ b_i ;
            Pn.Y = m_i ×P.Y + n_i ;
          }
          Bt.add(Pn);
}
...
```

Table 2. Each triangle (Ti ∈T) is enumerated to get the new POI set Bt

3.3. Automatic calculation of co-referential object

After position-correction in *Section 3.2*, One-Side Nearest Neighbor Method is used to get the co-referential objects that were omitted in Section 3.1. Giving a threshold to each point Ai in Set A, some points in Set Bt will be selected out as co-referential candidates.

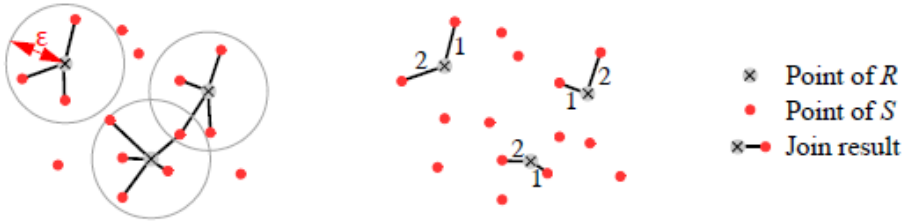

Figure 5. Use one-side nearest neighbor method to calculate co-referential object.

According to combination relation between name matching and address matching in *Table 1*, we can easily make sure that the co-referential object e is only in object Sets of M12,M21,M22,yet there are no same POI objects in M12,M22,M31,M32,M33, they are could not be co-referential objects. Therefore, co-referential object can be only in M12, M21, M22.

When select tactics of co-referential objects, this paper define that priority selection sequence of co-referential object in M12, M21, M22 is: M12 > M21 >M22. The reason is the granularity of user address labels are often not the same, making the probability of same address accuracy and completely equation so small, when address match is compatible and name match is exactly, objects to select M12 concentrated object priority.M21 may have different objects with same address, such as there are different POI objects in a building, objects address is exactly the same, but does not belong to the same object. The probability of this situation in the real world is relatively large, co-referential object probability in M21 is less than M12, but greater than M22.

4. Experiment and discussion

4.1. Experimental data and test results

In this paper, we manually collected some POI objects in the region of Lianhua Bridge in Haidian District of Beijing, China (W116.304997, S39.888172, E116.335521, N39.905651) on the map of www.mapbar.com and map.baidu.com. There are 219 POIs from MapABC (left one of *figure 6)* and 318 POIs from baidu (right one of *figure* 6).

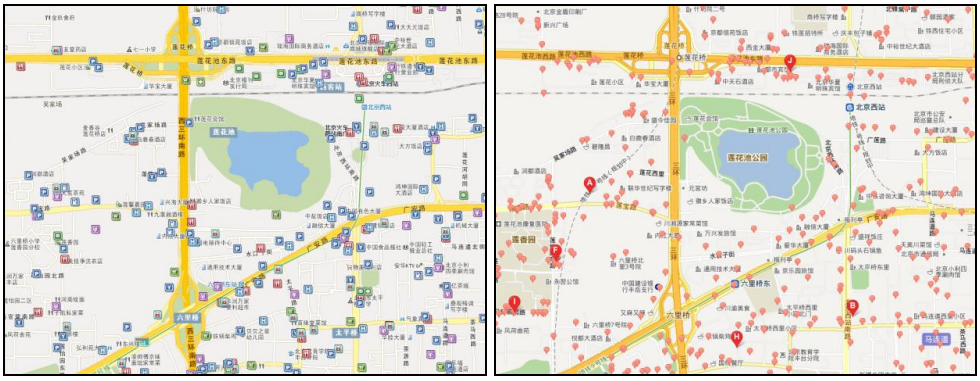

Figure 6. POI distribution map in experimental zone (left: mapbar, right: baidu).

First we manually checked each POI of the two datasets and get the number of co-referential entities (the total number of co-referential entities is 143).

Next simple text matching algorithm and name-address standardization algorithm were used respectively to find entities with same name and same address. Result of this step showed that we get 25 objects by using name - address standardized method, while simple text matching method only found 7.

After that Nearest Neighbor Method were used to find the rest of co-referential objects. First locally-position-correction based on triangulation and globally-position-correction were applied to POI from baidu.com, and we got two sets of "position-corrected" POIs; Then we use Nearest Neighbor Method (Threshold=10 meters) to the raw ses and the 2 "position-corrected" sets to calculate co-referential object. The results were shown in *Table 3*.

Position-correction	return results	correct results	recall rate	accuracy rate
None	54	32	27.1%	59.3%
globally-position	68	37	31.4%	54.4%
locally-position-correction based on triangulation	102	83	70.3%	81.4%

Table 3. Nearest Neighbor Method calculation results.

By summarizing results of all steps, we got *Table 4*. It shows that the method in this paper, by which the recall rate and accuracy rate respectively reach to 75.5% and 85.0%, has best effect

Position-correction	total number of co-referential objects	total number of correct co-referential objects	recall rate	accuracy rate
None	79	57	39.9%	72.2%
globally-position	93	62	43.4%	66.7%
locally-position-correction based on triangulation	127	108	75.5%	85.0%

Table 4. Algorithmic effect.

4.2. Analysis and Discussion

This paper proposed the method based on position-correction and semantic matching can effectively reduce the POI location deviation from different sources, and standardize the name, address and other information in POI. Experiment shows that it achieves effective in recall and precision rates and is a good method to deal with multi-source POI for finding the co-referential entities. It can be applied in the field of POI data fusion from different geographic information website particularly.

To further enhance the proposed method, it can be done in the following steps: it is necessary to establish Stable model of triangulation, do more research of calculating and merging the attribute of POI entities, and enhance the effectiveness and performance of mass data processing and so on.

This work was supported by the National High Technology Research & Development Program of China ("863" Program)(No. 2012AA12A402) , Science & Technology Development Plan of National Administration of Surveying, Mapping and Geoinformation (No. A11117) and the basic scientific research fund of CASM (No.7771209).

References

TIGER/LineFiles Technica1 Documentation. (2000) U.S. Department of Commerce, Geography Division U.S. CensusBureau

Böhm C, Krebs F (2002) High Performance Data Mining Using the Nearest Neighbor Join, Proceedings 2002 IEEE International Conference on Data Mining. ICDM 2002, P43 – 50

Beeri C, Kanza Y, Safra E, Sagiv E (2004) Object Fusion in Geographic Information Systems, VLDB '04 Proceedings of the Thirtieth international conference on Very large data bases- Volume 30, Pages 816 – 827

Li L, Goodchild MF (2010) Automatically and accurately matching objects in geo-spatial datasets. Theory, Data Handling and Modelling in GeoSpatial Information Science. Hong Kong, 26-28 May, 2010

Li L, Xu X, Wu Z, et. (2008) Research on Integration and Updating Method of Multi-sources Electronic Map Data. Geomatics and Information Science of Wuhan University, 33(4)

Peng Y, Peng z (2007) Research on Spatial Data Fusion Techniques. Computer Engineering, 33(18)

Liu X (2011) Study Segmentaion and Matching of Chinese POIs Based on Lucene. Computer Knowledge and Technology, 21(7):1009-3044.

Zhang X, Wang L (1997) Identification and Analysis of Chinese Organization and Institution Names. Journal of Chinese Information Processing, 4(11):p21-32

http://www.geog.buffalo.edu/ncgia/gishist/DIME_story.html

Li J, Wang D, Wang X (2008) Chinese organization name recognition based on template matching. Information Technology, 6(25):p97-99

Yu H, Zhang H, Liu Q (2003) Recognition of Chinese Organization Name Based on Role Tagging. Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, 2003